Informedia

Summarization-on-Demand

Auto-Summarization and Visualization Over Multiple Documents and Libraries

Alex Hauptmann

Carnegie Mellon University

Carnegie Mellon

Outline

- Visualization & Summarization Goals
- Visage Dynamic Query Histograms
- Informedia Result Set Visualization (Demo)
- Result Set Clustering and Labeling
- Summarization through Title Generation

Summarization and Visualization Over Multiple Documents

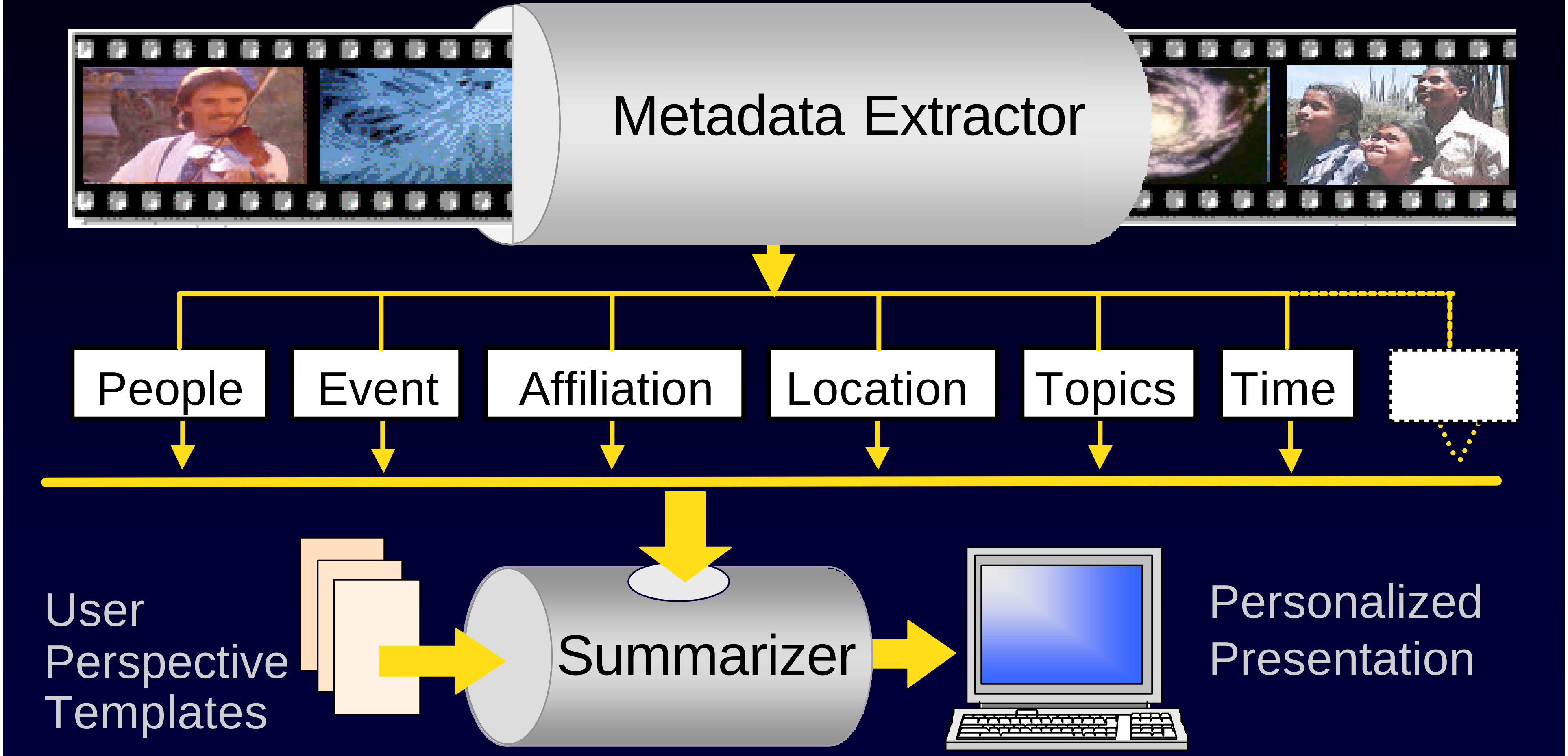
Goal: Auto-summarization and visualization of multiple audio & video documents, stories and reports

- *collages*: by time, location, personalities, events
- eliminating redundancy
- at any level of detail
- creating personality and event profiles on-demand

Requisite subgoals:

- establishing *geographic* and *temporal context*
  - geocoding all data by source and internal reference
  - extracting explicit and indirect time/date/place references

Metadata for Video Summarization



Carnegie
Mellon

Dynamic Query Histograms

Mark A. Derthick

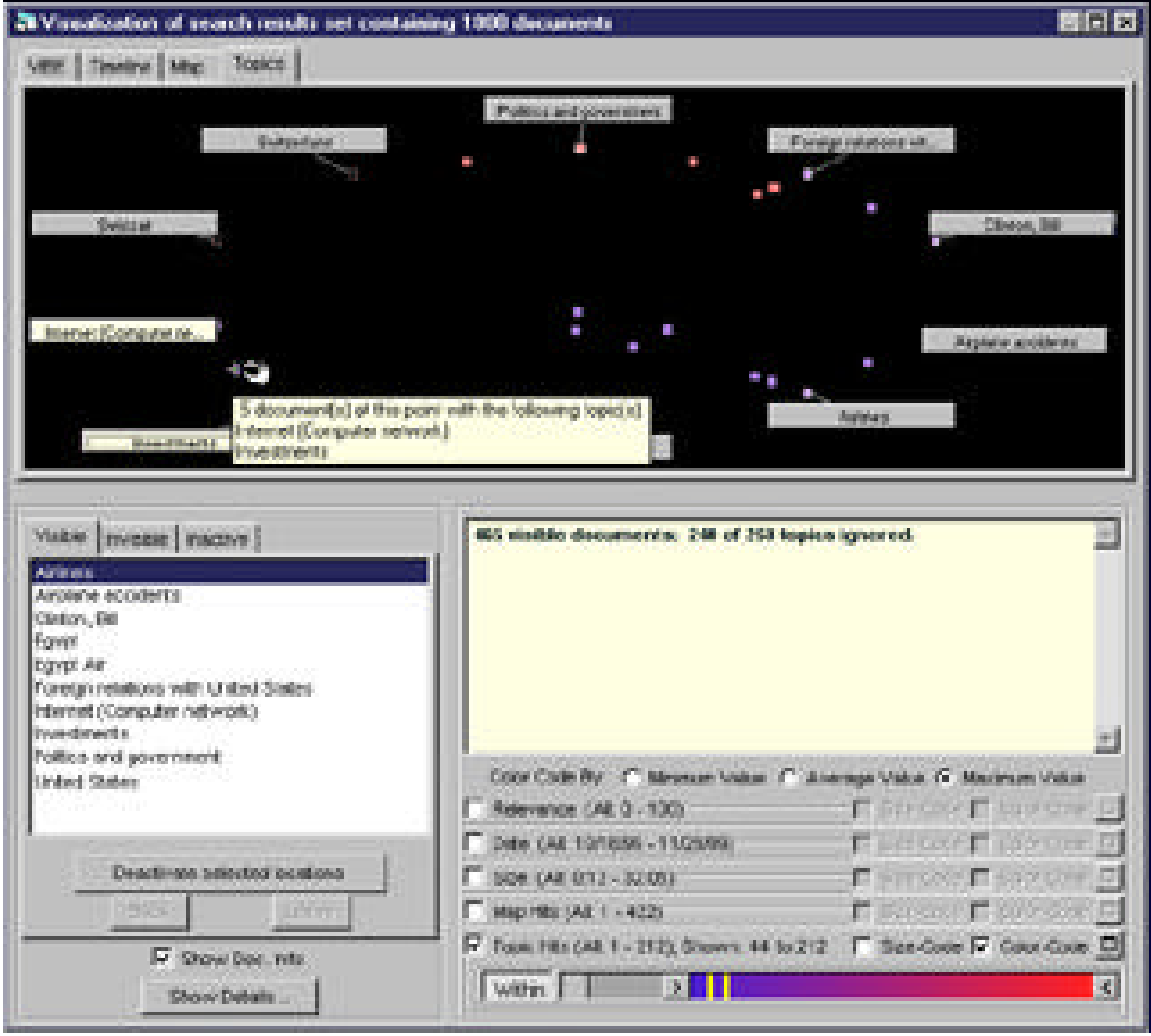
- Informedia library information through VISAGE data visualization tool
- Dynamically linking/updating different views on the same objects
- Efficient for large amounts of data

[redacted]

DLI-1: KNN-based Topic Detection

- Build training index with pre-labeled topics
  - 45000 Broadcast News stories from 1995 and 1996
  - 3178 different news topics occurring > 10 times
- Search for top 10 related stories in training index
- Lookup topics for related stories
- Re-weight topics by story relevance (select top 5)
- At 5 topics,   Recall - .491   Relevance - .482

Visualization of Result Space



Visualization of search results set containing 1000 documents

5 document(s) at this point with the following topic(s):
Internet (Computer network)
Investments

- Airlines
- Airplane accidents
- Clinton, Bill
- Egypt
- Egypt Air
- Foreign relations with United States
- Internet (Computer network)
- Investments
- Politics and government
- United States

865 visible documents. 248 of 258 topics ignored.

Color Code By: ○ Minimum Value ○ Average Value ● Maximum Value

□ Relevance (All 0 - 130) □ Size-Code □ Color-Code

□ Date (All 12/18/96 - 11/29/99) □ Size-Code □ Color-Code

□ Size (All 0.13 - 32.05) □ Size-Code □ Color-Code

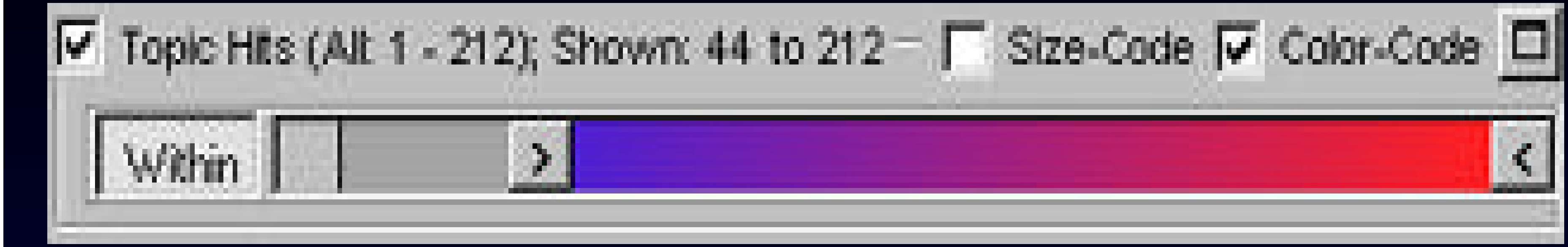
□ Map Hits (All 1 - 422) □ Size-Code □ Color-Code

☑ Topic Hits (All 1 - 212), Shown: 44 to 212 □ Size-Code ☑ Color-Code

☑ Show Doc. Info

Topic digest for recent news, color-coded and limited to frequently occurring topics

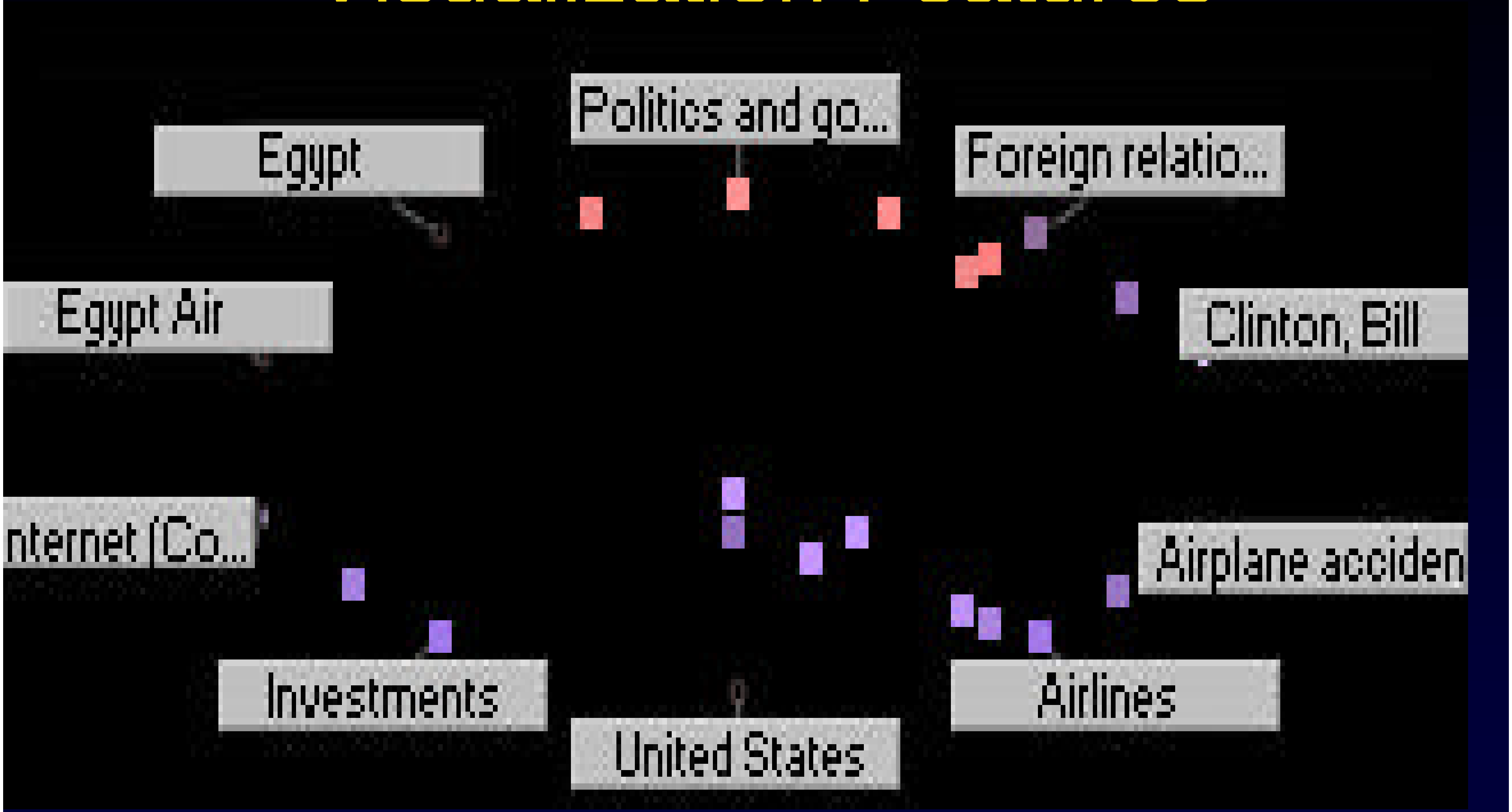
Visualization Features



☑ Topic Hits (All 1 - 212); Shown: 44 to 212 □ Size-Code ☑ Color-Code

Slider to reduce topic digest to frequently occurring topics

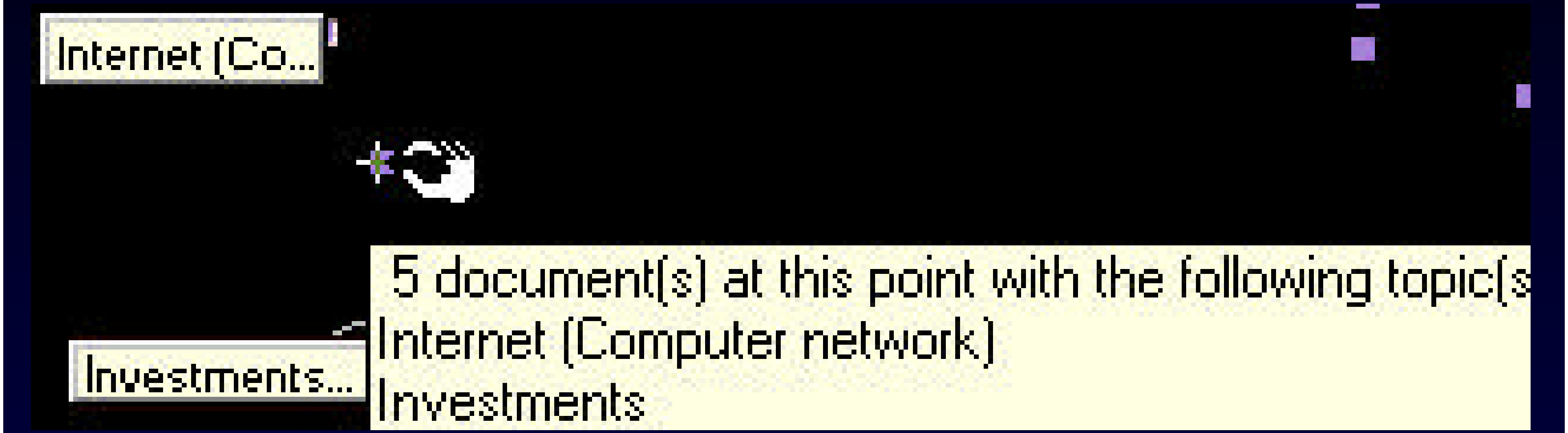
Visualization Features



Red color shows frequent topics and blue less frequent ones

Carnegie Mellon

Visualization Features

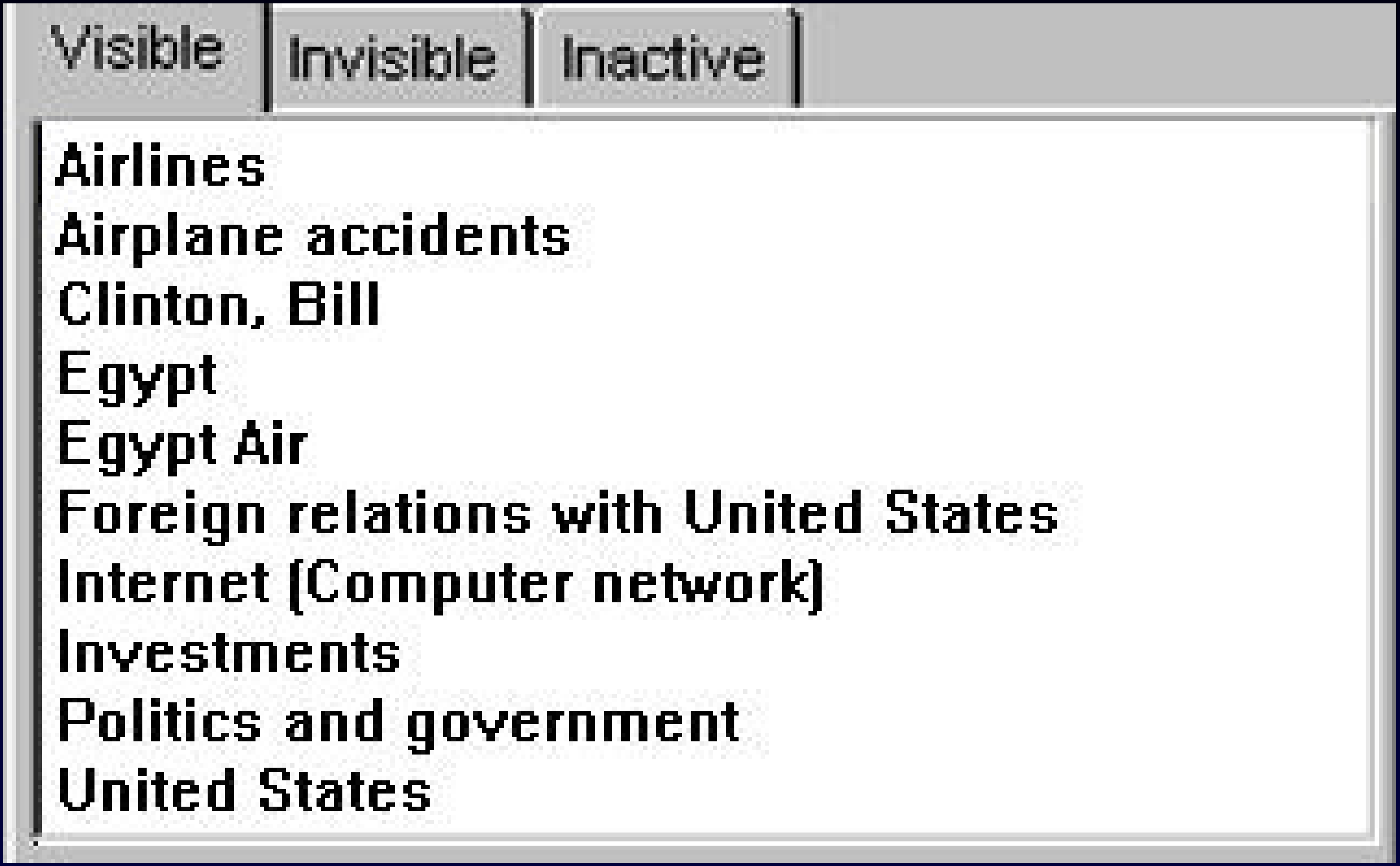


5 document(s) at this point with the following topic(s

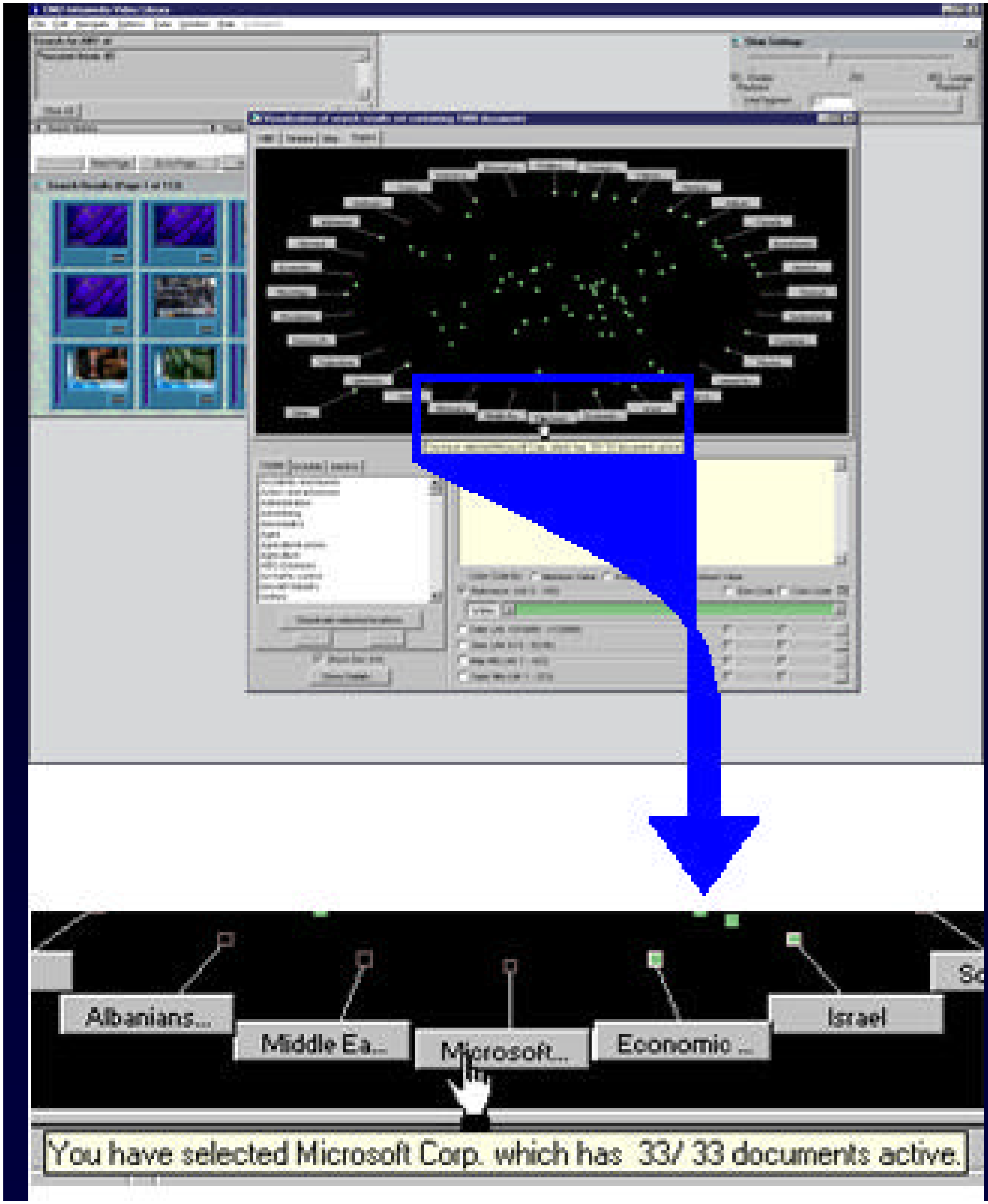
Internet (Computer network)

Investments

List of Active Topics



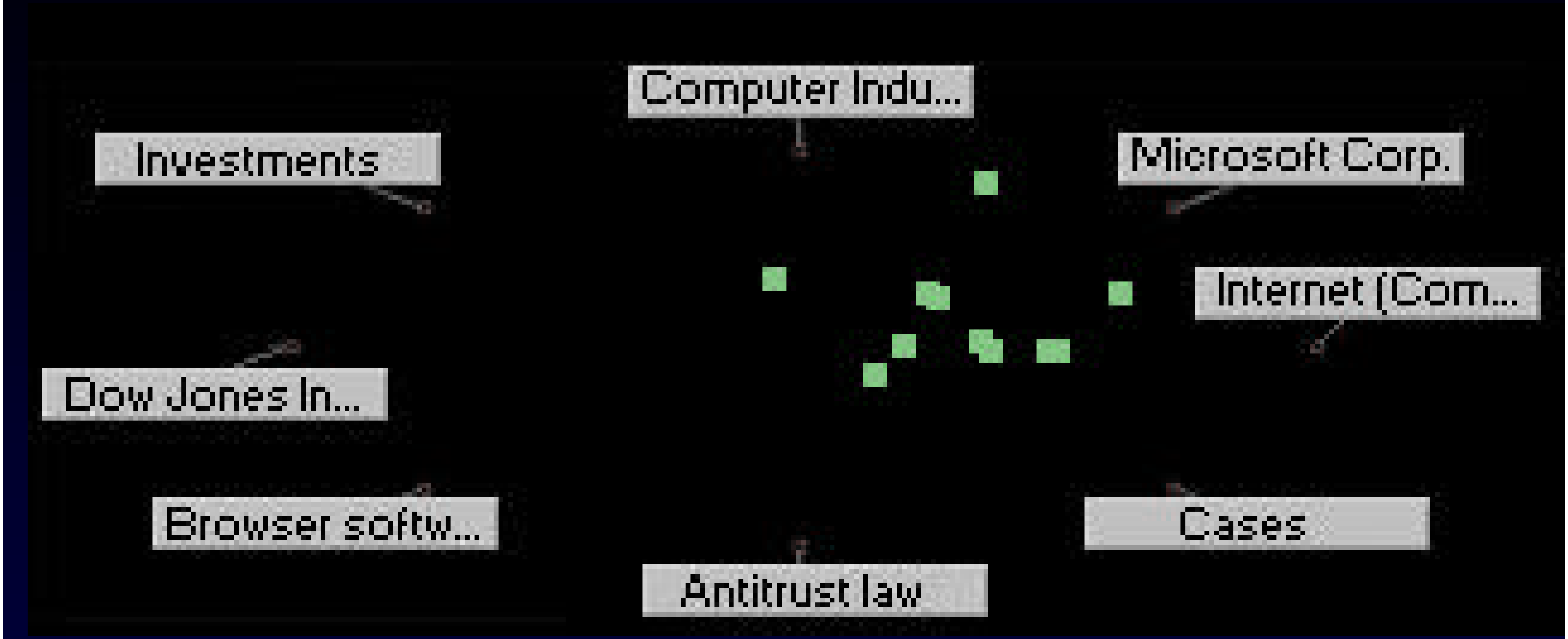
Airlines
Airplane accidents
Clinton, Bill
Egypt
Egypt Air
Foreign relations with United States
Internet (Computer network)
Investments
Politics and government
United States



You have selected Microsoft Corp. which has 33/ 33 documents active.

Zooming into documents with "Microsoft Corp." as a topic

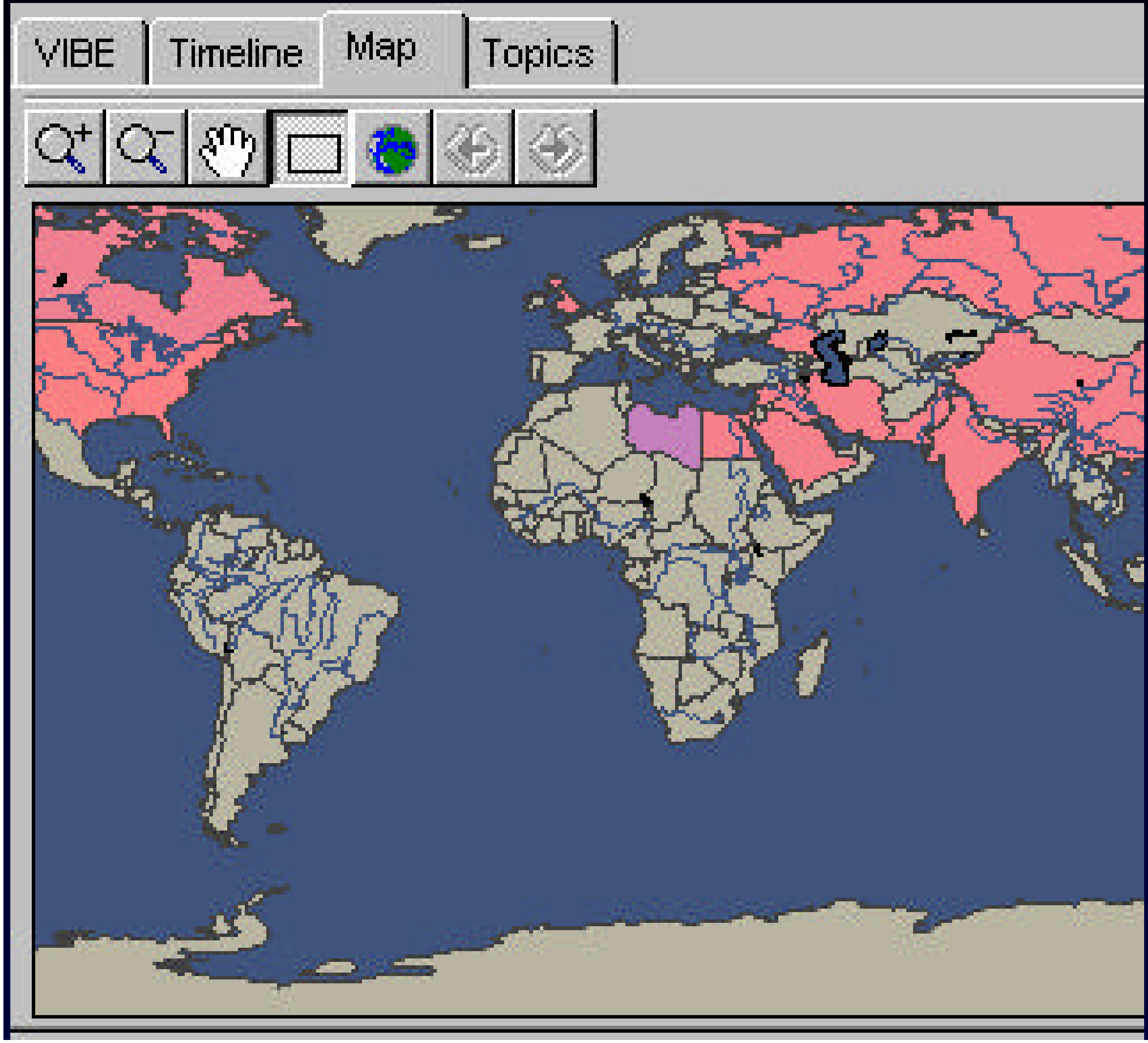
Visualization Features



Plot area after user zooms to documents labeled with "Microsoft Corp." topic

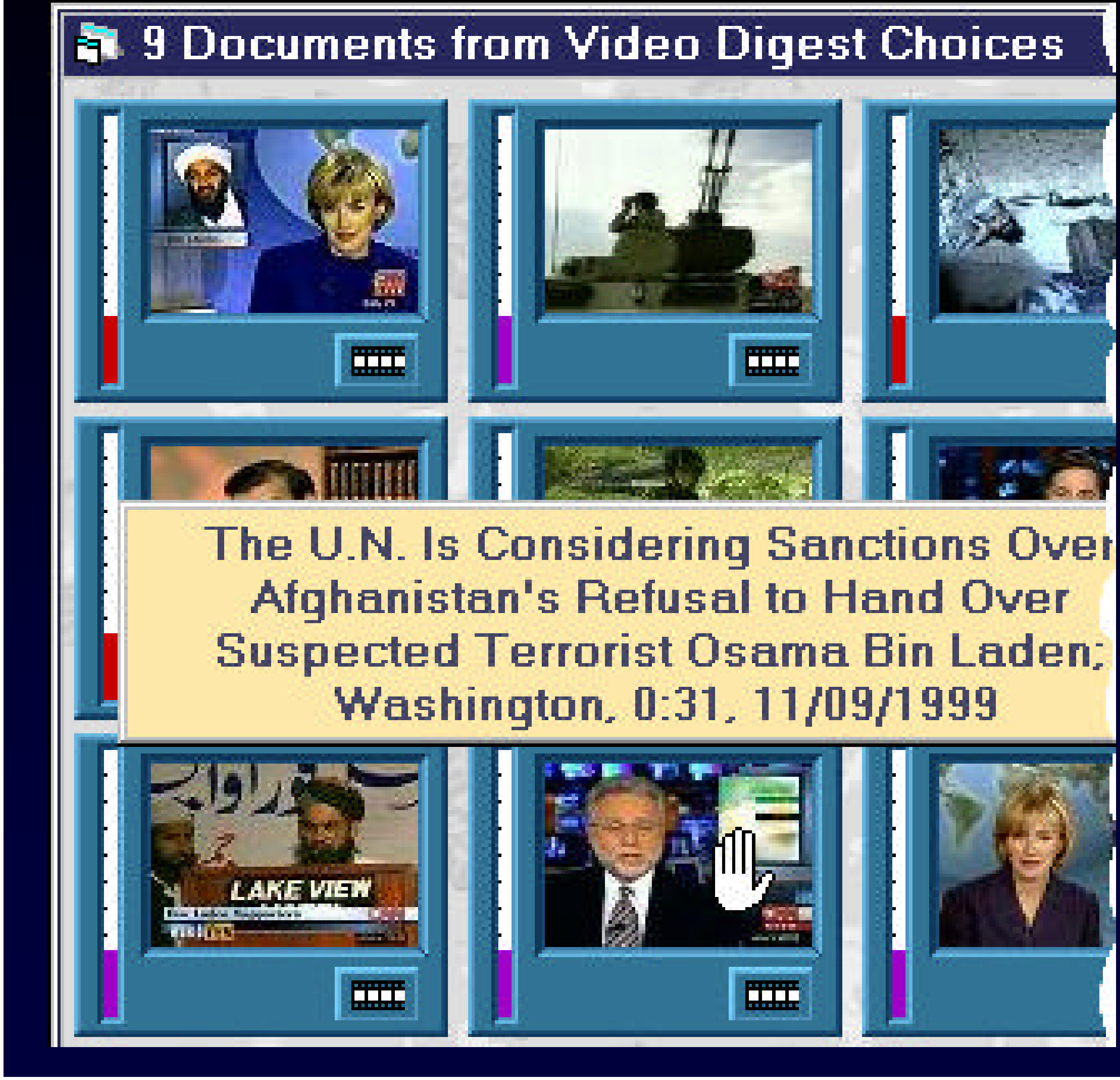
Carnegie Mellon

Map Digest View



color-coded to broadcast date (blue is older, red is newer) showing distribution of documents with “guided missiles” topic

Visualization Features

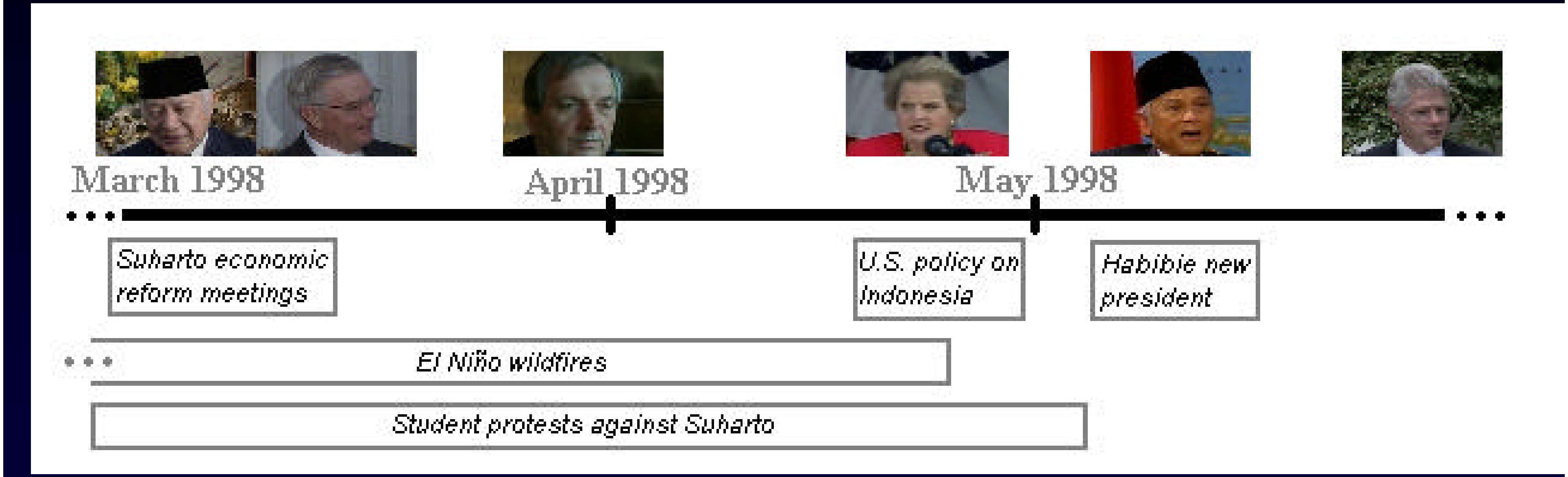

9 Documents from Video Digest Choices

The U.N. Is Considering Sanctions Over Afghanistan's Refusal to Hand Over Suspected Terrorist Osama Bin Laden; Washington, 0:31, 11/09/1999

View into small set of documents, showing thumbnail images and one title

Carnegie Mellon

Next Step: Topic/Face/Chrono-Collage



Title Generation for Informedia News Stories

- Informedia, a multimedia digital library, stores television broadcast news stories.
- An extractive summary feature currently locates snippets in news-story transcripts to use as story titles.
- GOAL: An improved, non-extractive title-generation feature for Informedia.

Basic Idea

- Train a statistical model on a corpus of documents with human-assigned titles.
- Compare 3 title generation methods:
  - Extractive Titles
  - Naïve Bayes
  - KNN
- Apply to machine translated documents

$$F1 = \frac{(2 * \text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

- Only measured word selection, not order

Experiment

- 40000 TV news stories with titles from 1998 Broadcast News CD-ROM
- tested on 10000 held-out stories evaluated on titles

Extractive Summarization

- MS Word 2000 AutoSummarize
- Extracts sentences/fragments as summaries
- Similar performance to TF IDF implementation at CMU
- Does not use our training corpus

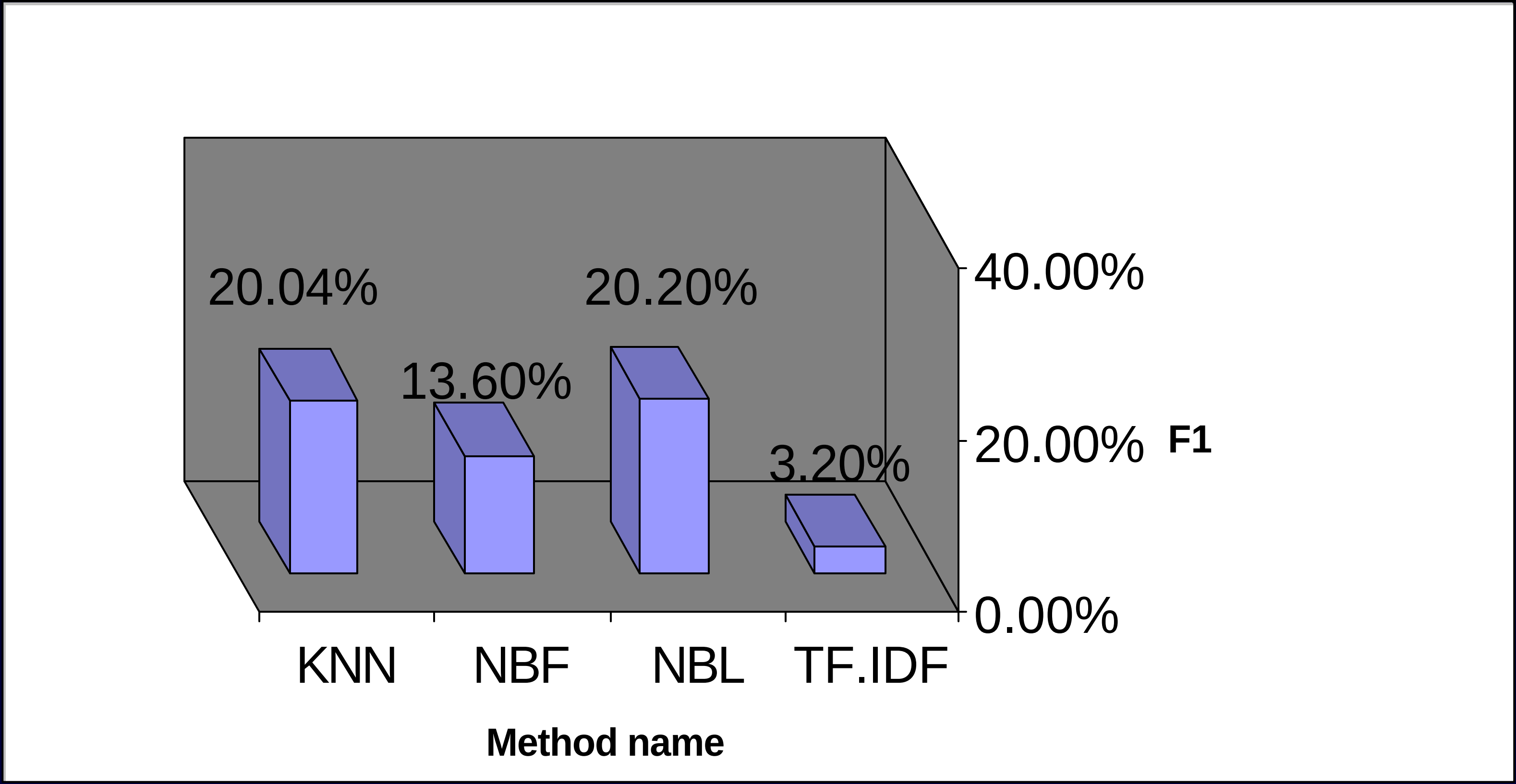
Naïve Bayes

- Train a statistical model on a corpus of documents with human-assigned titles.
- Title need not be a snippet from the document (contrasts with extractive-summarization techniques).
- Suggested by Witbrock & Mittal, 1999.
- $P(w_{Title}|w_{Doc})$
  - works better if $W_{title} = W_{Doc}$

(K) Nearest Neighbor

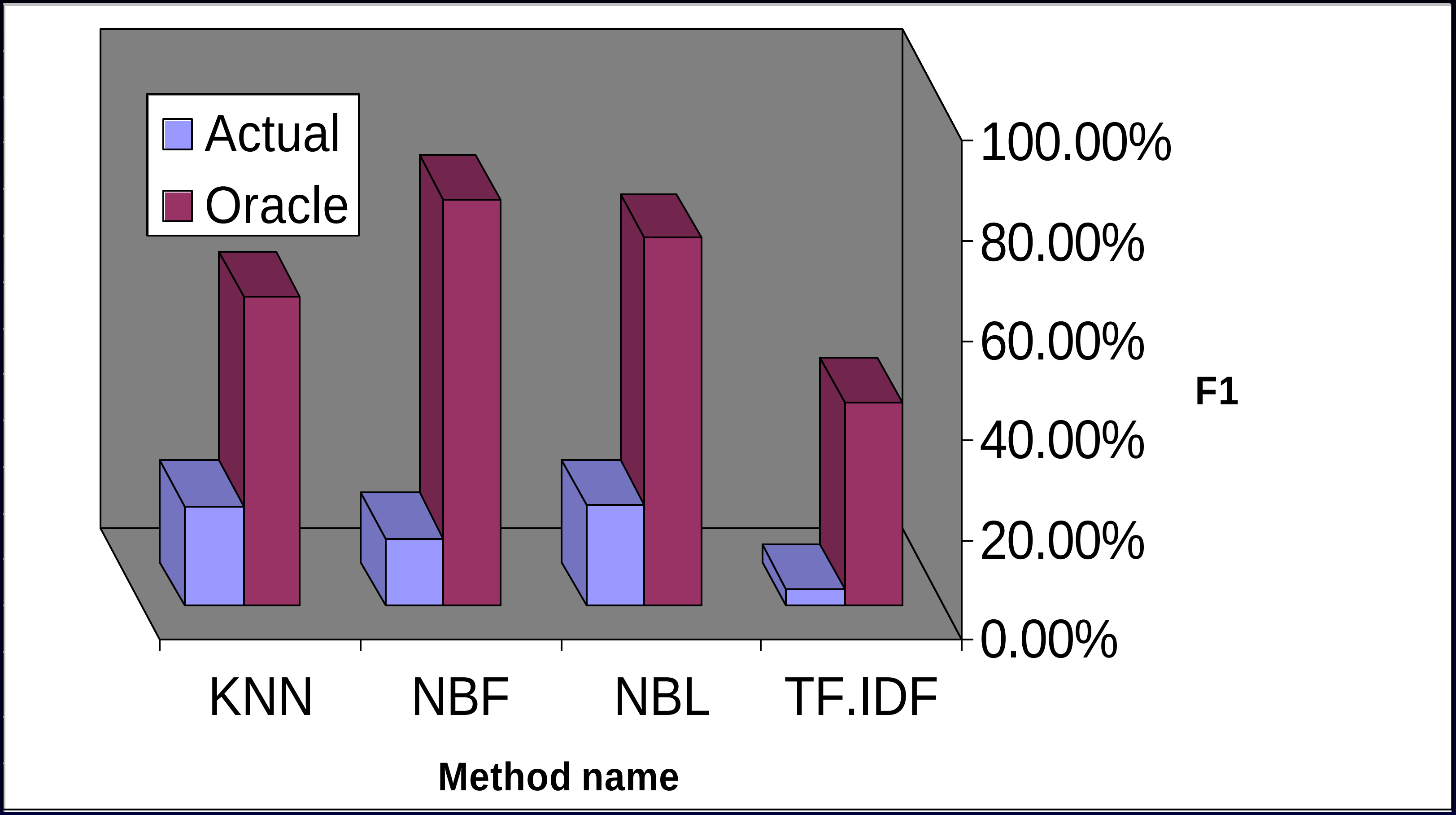
- Index a corpus of documents with human-assigned titles.
- Find the document in the training corpus closest to the current document
- Use that title (k=1)

Results



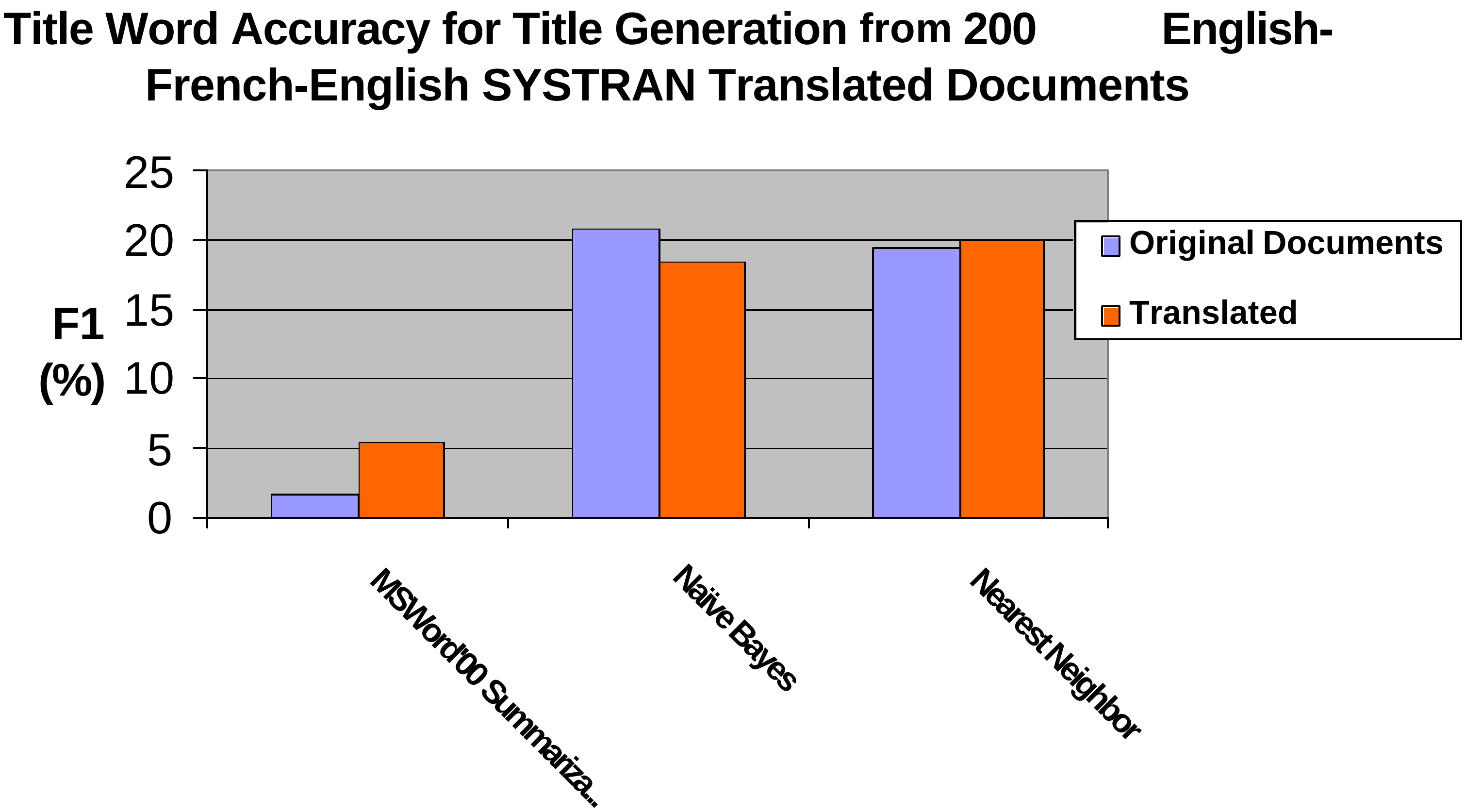
Carnegie Mellon

Simulated Best Possible Results



Carnegie Mellon

Translation Results



Carnegie Mellon

Orig Title: *INTERVIEW WITH CHINESE DISSIDENT WEI JINGSHENG*

KNN Title: *COMPLETE INTERVIEW WITH WEI JINGSHENG*

MSWORD AUTOSUM TITLE: *IT WAS A DECISION TO COME TO THE UNITED STATES ON BEHALF OF THE CHINESE GOVERNMENT STEP A DECISION TO SEEK THE GOOD PROMPT MEDICAL PROCESSING FOR YOU.*

Naïve Bayesian Title: *china wei chinese jingsheng china's prison*

SYSTRAN TRANSLATED TEXT (TO FRENCH AND BACK TO ENGLISH):

- IN AN EXCLUSIVE INTERVIEW WITH CHINESE DISSIDENT WEI OF C N N EXILED JINGSHENG INDICATES THAT IT SUPPORTS THE CONFRONTATION OF DIALOGUE NOT WITH THE CHINESE GOVERNMENT. RALPH BEGLEITER OF CORRESPONDENT OF BUSINESSES SAID BY WEI OF THE WORLD OF C N N HOW IT COULD COME TO THE UNITED STATES AND WHAT IT A DUE TO SUPPORT DURING ITS YEARS IN A CHINESE PRISON. IT WAS A DECISION TO COME TO THE UNITED STATES ON BEHALF OF THE CHINESE GOVERNMENT STEP A DECISION TO SEEK THE GOOD PROMPT MEDICAL PROCESSING FOR YOU. THEY CLEARLY INDICATED TO ME THAT IT IS IMPOSSIBLE SO THAT I REMAIN IN CHINA. IF I WANT TO OBTAIN A MEDICAL PROCESSING THAT YOU MUST GO ABROAD AND YOU MUST GO TO THE UNITED STATES. IT IS A CONDITION WHICH THEY CLEARLY INDICATED TO ME THUS. SAY A LITTLE TO ME ABOUT THE TRUTH OF YOUR PROCESSING IN PRISON. WERE YOU TORTURED WERE YOU WOUNDED BY THE GOVERNMENT OR OTHER PRISONERS. YES. USUALLY THESE BEATS ARE NOT DIRECTLY OF THE FONT. THEY USE USUALLY CRIMINALS WITH BEATUP OF OTHER CRIMINALS. IT BEATS ME NOT ONLY THAT IT IS COMMON OCCURRENCE WHICH TORTURES AND WHILE BEATING PRODUCE YOU IN THE PRISON. WEI INDICATES THAT IT SE RETURNS ACCOUNT WHICH IT THERE A OF A MORE ECONOMIC FREEDOM IN CHINA THAN THERE WAS BUT IT SAYS THAT POLITICAL FREEDOM IS MORE SIGNIFICANT AND THAT ONE REFUSED TO HIM WITH

  THE PEOPLE OF CHINA

Title Generation Conclusions

- Training on a corpus really helps
- But coverage of training to testing better be good
- KNN is preferred since it provides a good readable title
- KNN could be improved with K > 1,
  - but then we need to combine title fragments
- Title word learning can also be improved
  - but we need robust word serialization with a language model

Other Insights

- Similar performance achieved using news stories with titles harvested from the web
- EM instead of Naïve Bayes is slightly better
- Small degradation for corrupted text (speech recognition, machine translation)

Summary of Summaries and Visualizations

- Exploring different paradigms of visualization
  - No ‘Right Way’

- Summarization through labeling and titles

The End